

Assessment: Exploring the Gapminder Dataset

#####Exercise 1. Life expectancy vs fertility - part 1 The Gapminder Foundation (www.gapminder.org) is a non-profit organization based in Sweden that promotes global development through the use of statistics that can help reduce misconceptions about global development.

Instructions 100 XP Using ggplot and the points layer, create a scatter plot of life expectancy versus fertility for the African continent in 2012. Remember that you can use the R console to explore the gapminder dataset to figure out the names of the columns in the dataframe. In this exercise we provide parts of code to get you going. You need to fill out what is missing. But note that going forward, in the next exercises, you will be required to write most of the code.

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
## filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
## intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

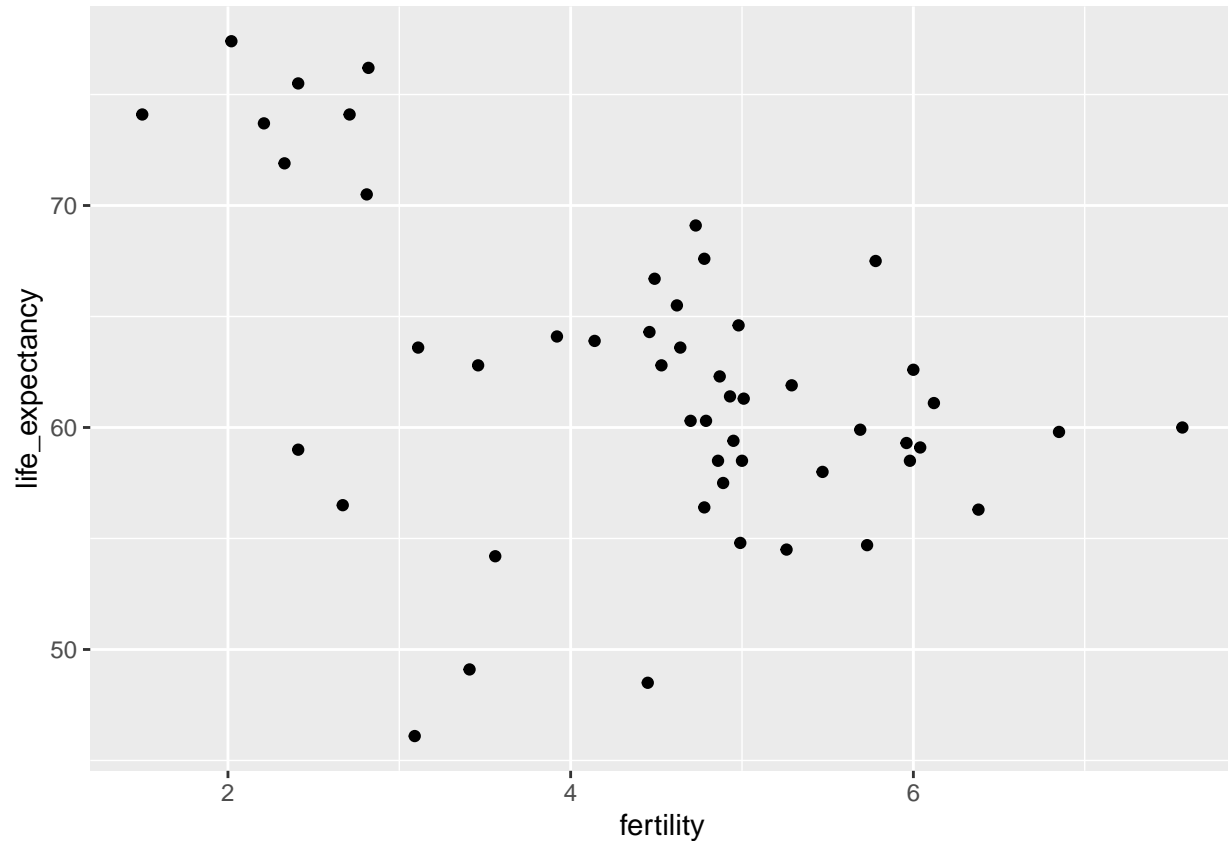
```
library(dslabs)
```

```
data(gapminder)
```

```
head(gapminder)
```

```
##           country year infant_mortality life_expectancy fertility
## 1      Albania 1960      115.40           62.87           6.19
## 2      Algeria 1960      148.20           47.50           7.65
## 3        Angola 1960      208.00           35.98           7.32
## 4 Antigua and Barbuda 1960      NA           62.97           4.43
## 5      Argentina 1960      59.87           65.39           3.11
## 6      Armenia 1960      NA           66.86           4.55
## population      gdp continent      region
## 1    1636054      NA    Europe Southern Europe
## 2    11124892 13828152297    Africa Northern Africa
## 3     5270844      NA    Africa  Middle Africa
## 4      54681      NA  Americas    Caribbean
## 5    20619075 108322326649  Americas    South America
## 6     1867396      NA     Asia   Western Asia
```

```
## fill out the missing parts in filter and aes
gapminder %>% filter(continent=="Africa" & year=="2012" ) %>%
  ggplot(aes(y=life_expectancy,x=fertility)) +
  geom_point()
```



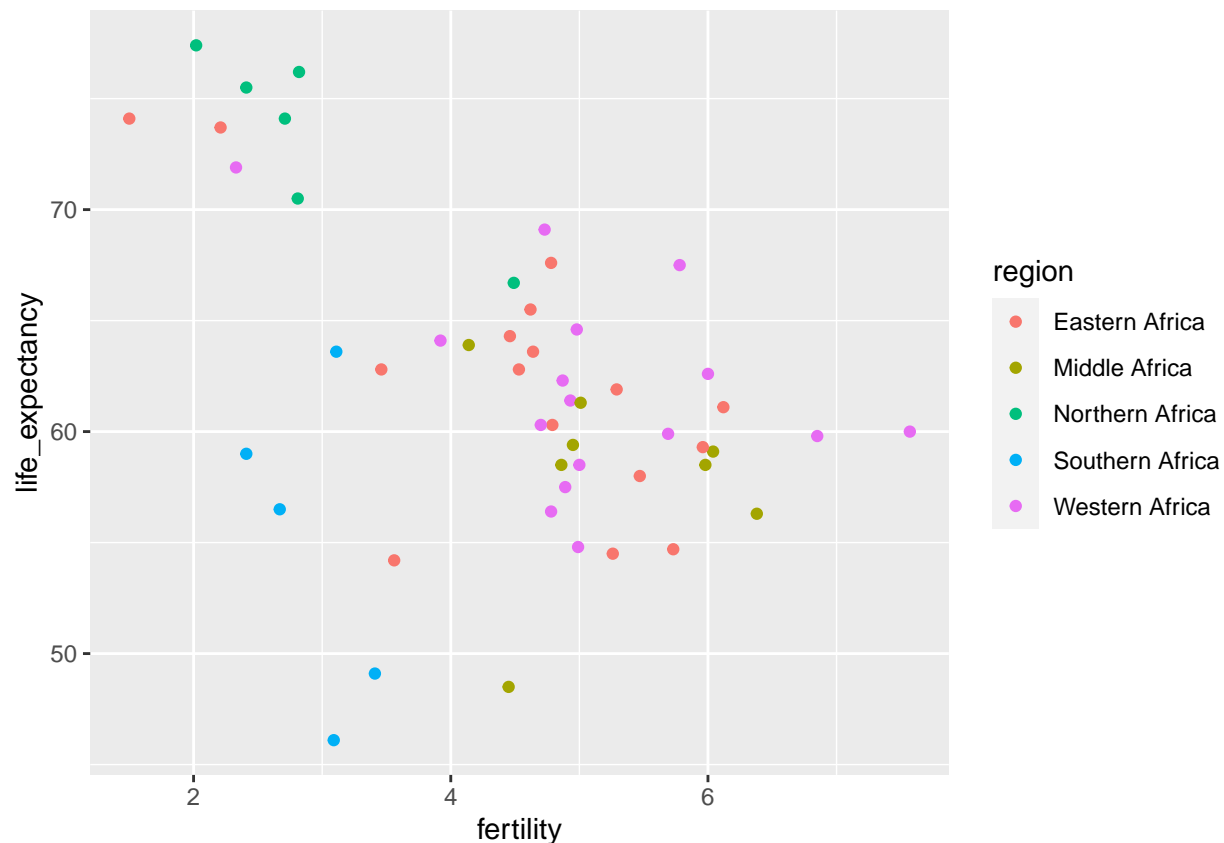
Exercise 2. Life expectancy vs fertility - part 2 - coloring your plot Note that there is quite a bit of variability in life expectancy and fertility with some African countries having very high life expectancies. There also appear to be three clusters in the plot.

Instructions 100 XP Remake the plot from the previous exercises but this time use color to distinguish the different regions of Africa to see if this explains the clusters. Remember that you can explore the gapminder data to see how the regions of Africa are labeled in the data frame!

Use color rather than col inside your ggplot call - while these two forms are equivalent in R, the grader specifically looks for color.

```
library(dplyr)
library(ggplot2)
library(dslabs)
data(gapminder)

gapminder %>% filter(continent=="Africa" & year=="2012" ) %>%
  ggplot(aes(y=life_expectancy,x=fertility,color=region)) +
  geom_point()
```



Exercise 3. Life expectancy vs fertility - part 3 - selecting country and region While many of the countries in the high life expectancy/low fertility cluster are from Northern Africa, three countries are not.

Instructions 100 XP Create a table showing the country and region for the African countries (use select) that in 2012 had fertility rates of 3 or less and life expectancies of at least 70. Assign your result to a data frame called df.

```
library(dplyr)
library(dslabs)
data(gapminder)
```

```
df <- gapminder %>% filter(continent=="Africa" & year=="2012" & fertility<=3 & life_expectancy>=70) %>%
```

Exercise 4. Life expectancy and the Vietnam War - part 1 The Vietnam War lasted from 1955 to 1975. Do the data support war having a negative effect on life expectancy? We will create a time series plot that covers the period from 1960 to 2010 of life expectancy for Vietnam and the United States, using color to distinguish the two countries. In this start we start the analysis by generating a table.

Instructions 100 XP Use filter to create a table with data for the years from 1960 to 2010 in Vietnam and the United States. Save the table in an object called tab.

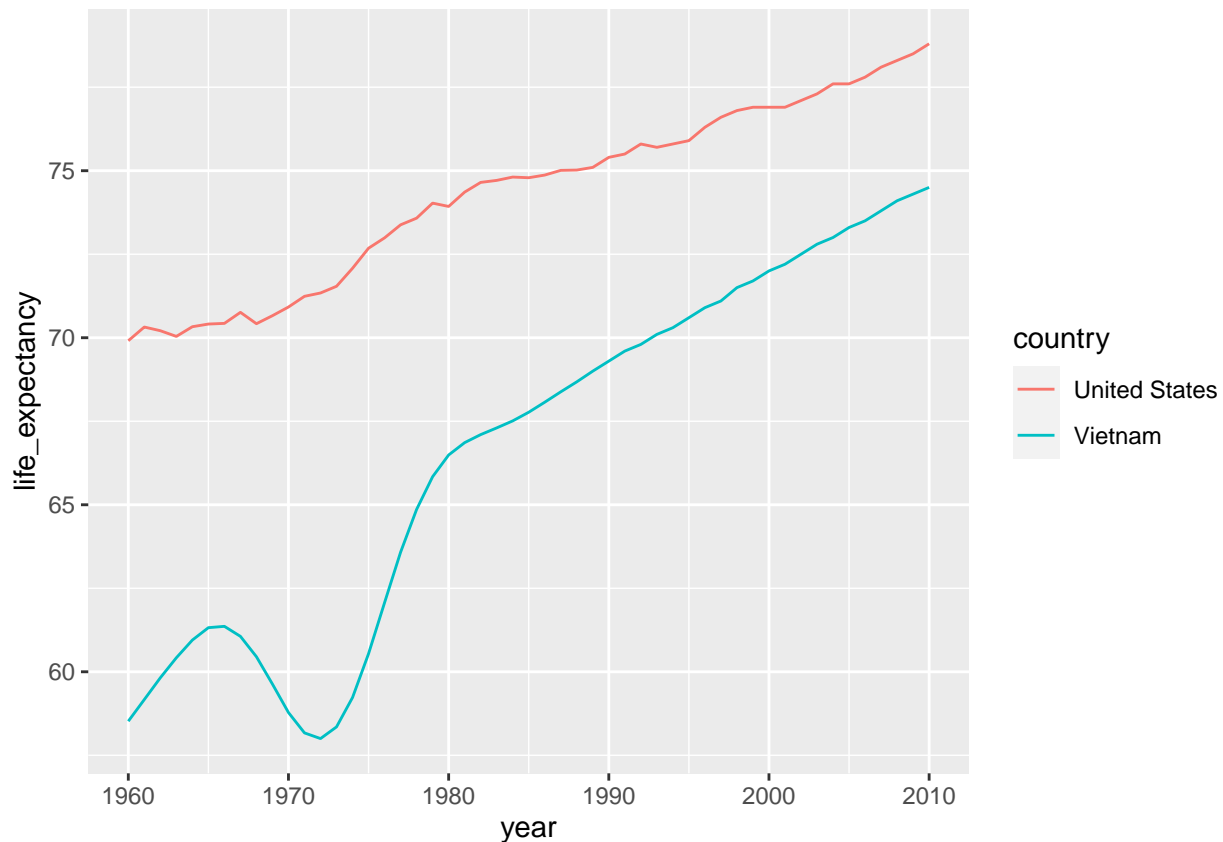
```
library(dplyr)
library(dslabs)
data(gapminder)
```

```
tab <- gapminder %>% filter(year>=1960 & year<=2010 & (country=="United States" | country=="Vietnam" ))
```

Exercise 5. Life expectancy and the Vietnam War - part 2 Now that you have created the data table in Exercise 4, it is time to plot the data for the two countries.

Instructions 100 XP Use `geom_line` to plot life expectancy vs year for Vietnam and the United States and save the plot as `p`. The data table is stored in `tab`. Use color to distinguish the two countries. Print the object `p`.

```
countries <- c("United States","Vietnam")
p <- tab %>% filter(country %in% countries) %>%
  ggplot(aes(x=year, y=life_expectancy, group = country,color=country)) +
  geom_line()
# code for your plot goes here - the data table is stored as `tab`
p
```



Exercise 6. Life expectancy in Cambodia Cambodia was also involved in this conflict and, after the war, Pol Pot and his communist Khmer Rouge took control and ruled Cambodia from 1975 to 1979. He is considered one of the most brutal dictators in history. Do the data support this claim?

Instructions 100 XP Use a single line of code to create a time series plot from 1960 to 2010 of life expectancy vs year for Cambodia.

```
library(dplyr)
library(ggplot2)
library(dslabs)
data(gapminder)

gapminder %>%
  filter(country == "Cambodia" & year>=1960 & year<=2010) %>%
  ggplot(aes(year, life_expectancy)) +
  geom_line()
```



Exercise 7. Dollars per day - part 1 Now we are going to calculate and plot dollars per day for African countries in 2010 using GDP data.

In the first part of this analysis, we will create the dollars per day variable.

Instructions 100 XP Use mutate to create a dollars_per_day variable, which is defined as $\text{gdp}/\text{population}/365$. Create the dollars_per_day variable for African countries for the year 2010. Remove any NA values. Save the mutated dataset as daydollars.

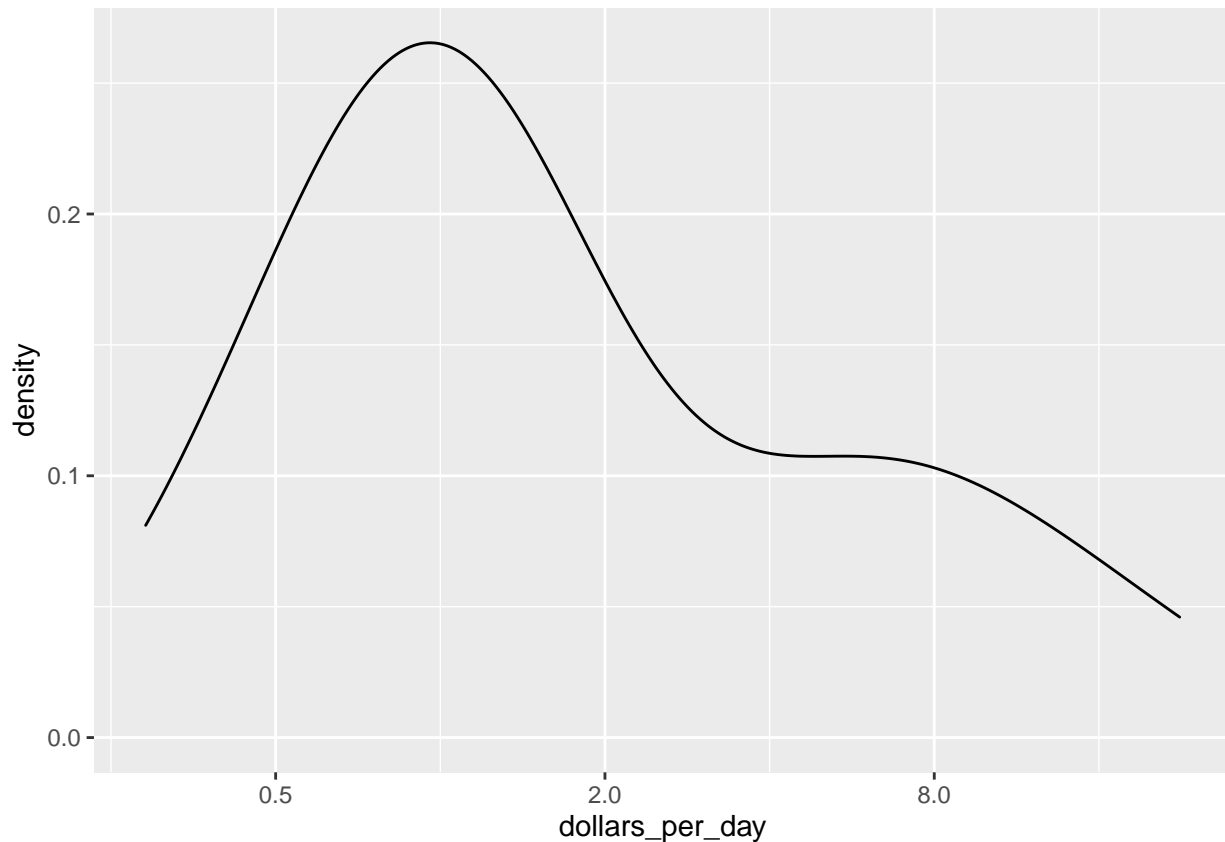
```
library(dplyr)
library(dslabs)
data(gapminder)
daydollars <- gapminder %>% mutate(., dollars_per_day=gdp/population/365)%>% filter(continent=="Africa")
```

Exercise 8. Dollars per day - part 2 Now we are going to calculate and plot dollars per day for African countries in 2010 using GDP data.

In the second part of this analysis, we will plot the smooth density plot using a log (base 2) x axis.

Instructions 100 XP The dataset including the dollars_per_day variable is preloaded as daydollars. Create a smooth density plot of dollars per day from daydollars. Use scale_x_continuous to change the x-axis to a log (base 2) scale.

```
# your code here
daydollars%>% ggplot(aes(dollars_per_day, y = )) +
  scale_x_continuous(trans = "log2")+ geom_density(alpha = 0.2, bw = 0.75)
```

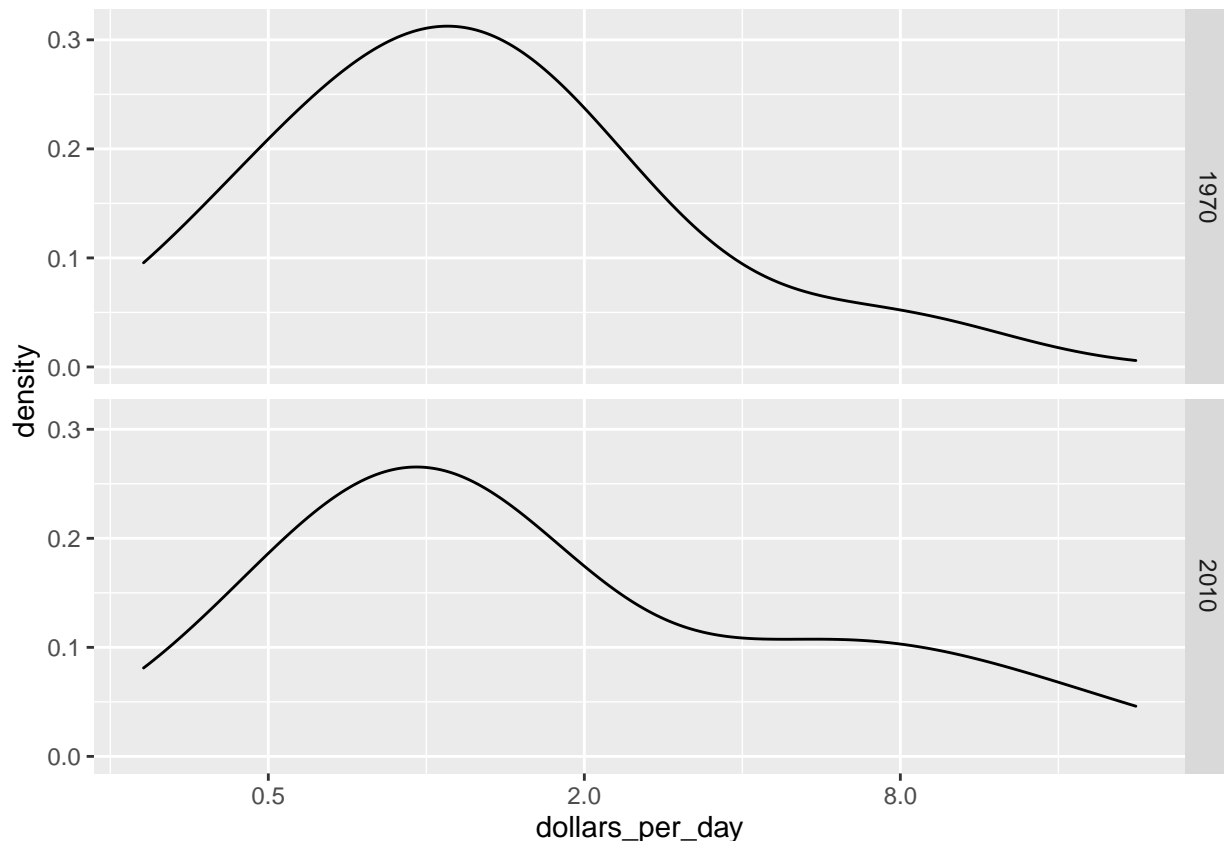


Exercise 9. Dollars per day - part 3 - multiple density plots Now we are going to combine the plotting tools we have used in the past two exercises to create density plots for multiple years.

Instructions 100 XP Create the dollars_per_day variable as in Exercise 7, but for African countries in the years 1970 and 2010 this time. Make sure you remove any NA values. Create a smooth density plot of dollars per day for 1970 and 2010 using a log (base 2) scale for the x axis. Use facet_grid to show a different density plot for 1970 and 2010.

```
library(dplyr)
library(ggplot2)
library(dslabs)
data(gapminder)
```

```
daydollars <- gapminder %>% mutate(.,dollars_per_day=gdp/population/365)%>% filter(continent=="Africa")
daydollars%>% ggplot(aes(dollars_per_day, y = )) +
  scale_x_continuous(trans = "log2")+ geom_density(alpha = 0.2, bw = 0.75)+ facet_grid(year ~ .)
```

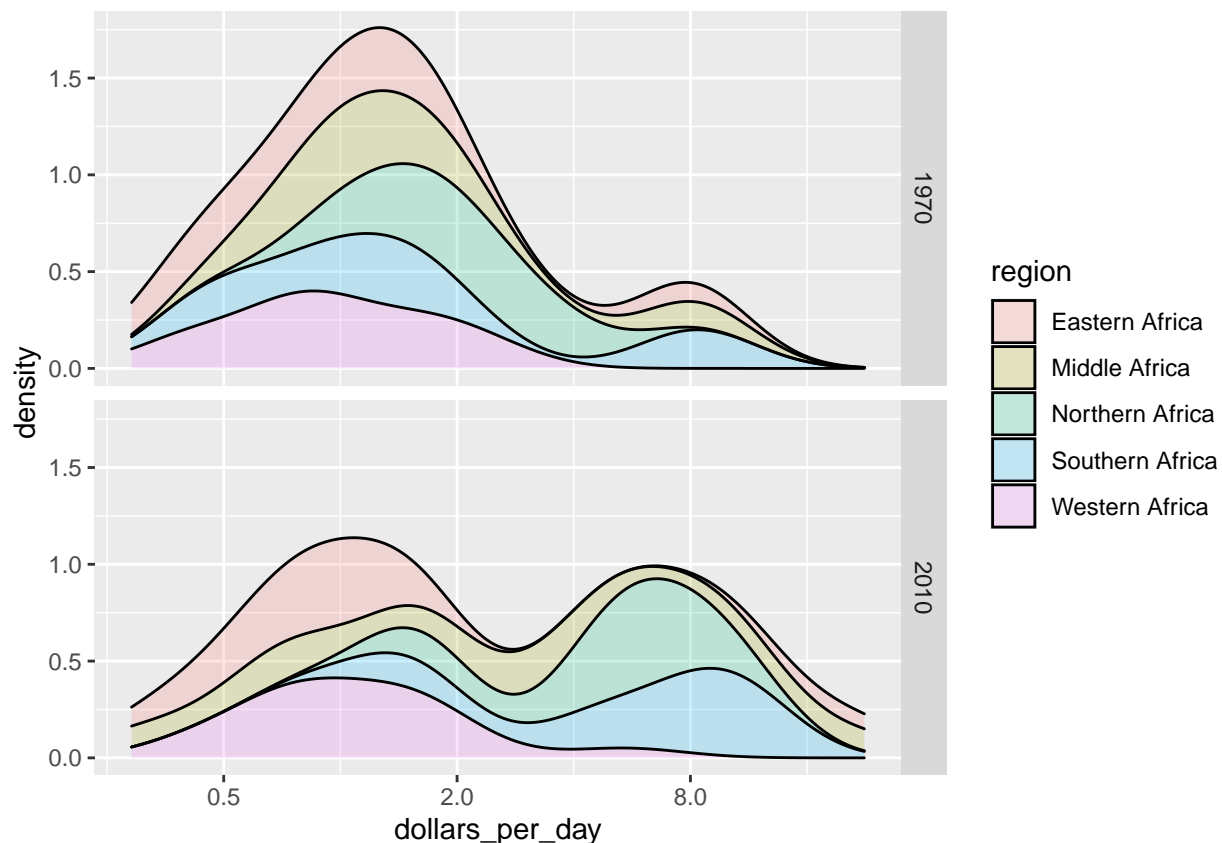


Exercise 10. Dollars per day - part 4 - stacked density plot Now we are going to edit the code from Exercise 9 to show a stacked density plot of each region in Africa.

Instructions 100 XP Much of the code will be the same as in Exercise 9: Create the `dollars_per_day` variable as in Exercise 7, but for African countries in the years 1970 and 2010 this time. Make sure you remove any NA values. Create a smooth density plot of dollars per day for 1970 and 2010 using a log (base 2) scale for the x axis. Use `facet_grid` to show a different density plot for 1970 and 2010. Make sure the densities are smooth by using `bw = 0.5`. Use the `fill` and `position` arguments where appropriate to create the stacked density plot of each region.

```
library(dplyr)
library(ggplot2)
library(dslabs)
data(gapminder)

gapminder %>% mutate(.,dollars_per_day=gdp/population/365)%>% filter(continent=="Africa" & year %in% c
```

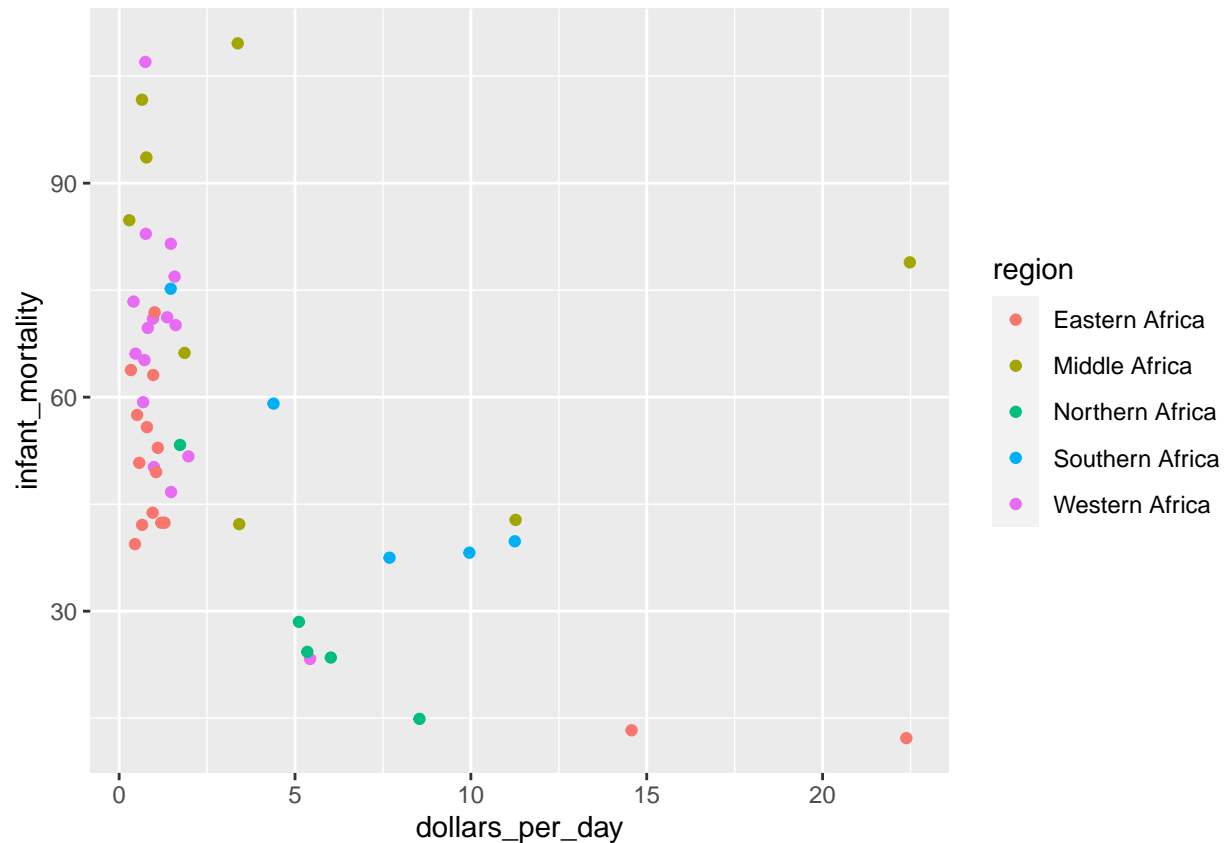


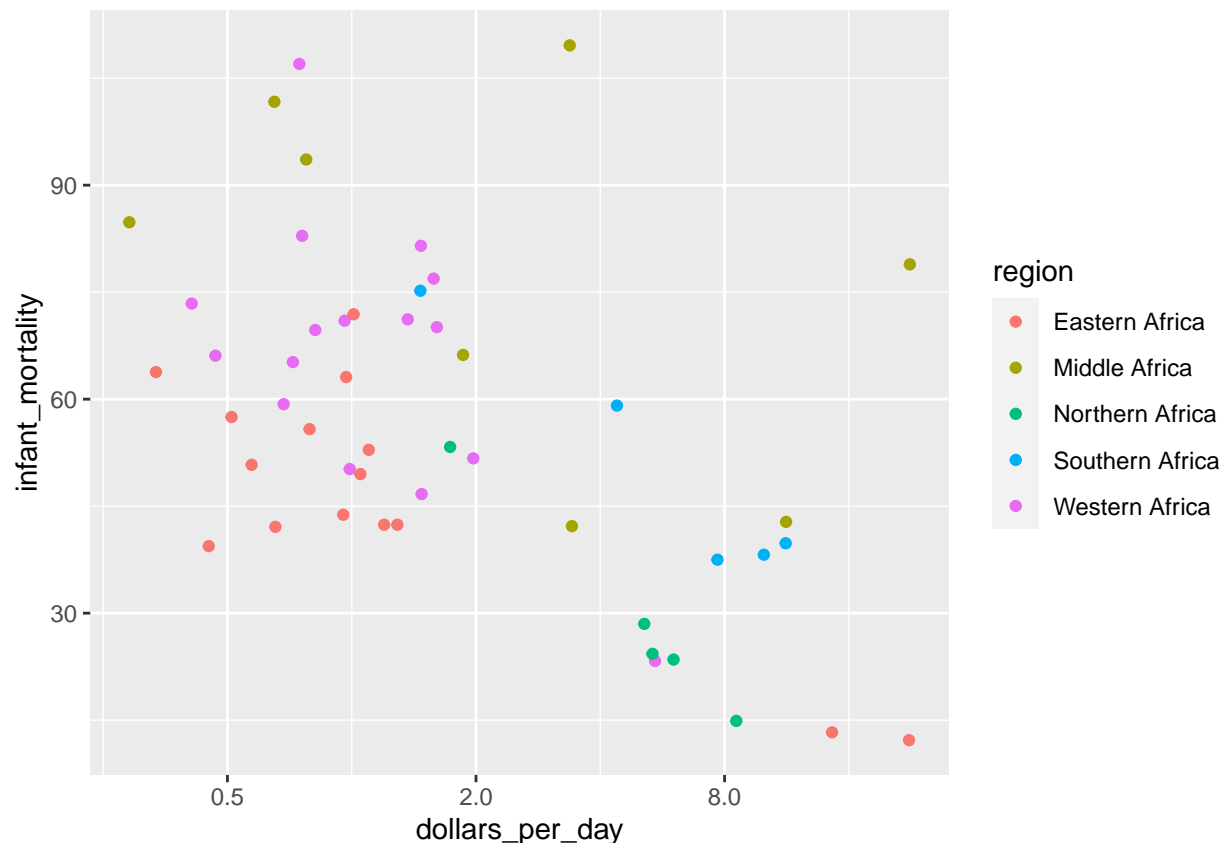
Exercise 11. Infant mortality scatter plot - part 1 We are going to continue looking at patterns in the gapminder dataset by plotting infant mortality rates versus dollars per day for African countries.

Instructions 100 XP Generate `dollars_per_day` using `mutate` and `filter` for the year 2010 for African countries. Remember to remove NA values. Store the mutated dataset in `gapminder_Africa_2010`. Make a scatter plot of `infant_mortality` versus `dollars_per_day` for countries in the African continent. Use color to denote the different regions of Africa.

```
library(dplyr)
library(ggplot2)
library(dslabs)
data(gapminder)
gapminder_Africa_2010 <- gapminder %>% mutate(.,dollars_per_day=gdp/population/365)%>% filter(continent == "Africa")

# now make the scatter plot
gapminder_Africa_2010 %>% ggplot(aes(x=dollars_per_day,y=infant_mortality, color = region)) +geom_point()
```



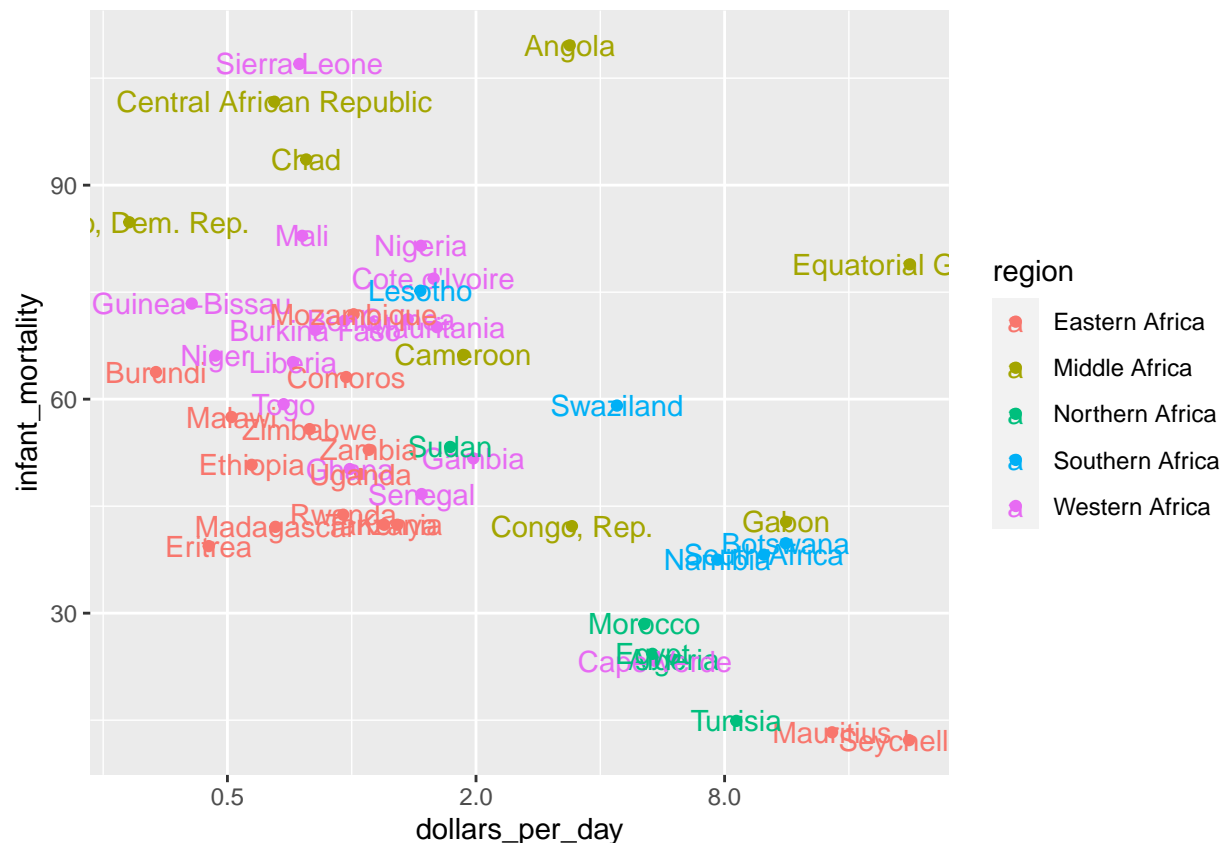
Exercise 13. Infant mortality scatter plot - part 3 - adding labels Note that there is a large variation in infant mortality and dollars per day among African countries.

As an example, one country has infant mortality rates of less than 20 per 1000 and dollars per day of 16, while another country has infant mortality rates over 10% and dollars per day of about 1.

In this exercise, we will remake the plot from Exercise 12 with country names instead of points so we can identify which countries are which.

Instructions 100 XP The mutated dataset is preloaded as `gapminder_Africa_2010`. As in the previous exercise, make a scatter plot of `infant_mortality` versus `dollars_per_day` for countries in the African continent. As in the previous exercise, use color to denote the different regions of Africa. As in the previous exercise, transform the x axis to be in the log (base 2) scale. Add a `geom_text` layer to display country names in addition to of points.

```
gapminder_Africa_2010 %>% ggplot(aes(x=dollars_per_day,y=infant_mortality, color = region,label=country))
```



Exercise 14. Infant mortality scatter plot - part 4 - comparison of scatter plots Now we are going to look at changes in the infant mortality and dollars per day patterns African countries between 1970 and 2010.

Instructions 100 XP Generate dollars_per_day using mutate and filter for the years 1970 and 2010 for African countries. Remember to remove NA values. As in the previous exercise, make a scatter plot of infant_mortality versus dollars_per_day for countries in the African continent. As in the previous exercise, use color to denote the different regions of Africa. As in the previous exercise, transform the x axis to be in the log (base 2) scale. As in the previous exercise, add a layer to display country names instead of points. Use facet_grid to show different plots for 1970 and 2010. Align the plots vertically.

```
library(dplyr)
library(ggplot2)
library(dslabs)
data(gapminder)
gapminder %>%
  mutate(dollars_per_day = gdp/population/365) %>%
  filter(continent == "Africa" & year %in% c(1970, 2010) & !is.na(dollars_per_day) & !is.na(infant_mortality))
ggplot(aes(dollars_per_day, infant_mortality, color = region, label = country)) +
  geom_text() +
  scale_x_continuous(trans = "log2") +
  facet_grid(year~.)
```

